shares imply is also what the article title implies is a convenient way to check that a topic model has succeeded in capturing important themes in a collection of texts.[23]

Four German Studies Journals (1928–2006)

To explore the corpus of journal articles using LDA, I fixed the number of topics at a hundred.[24] As described previously, LDA infers the distri- bution of the hundred topics across all the articles in the corpus as well as words characteristic of each topic. When we examine the inferred top- ics and plot their prevalence over the twentieth century, two dominant trends emerge. The first trend is a decline in articles on language peda- gogy. Topic 64 captures this trend neatly. Its characteristic words include "students," "language," "course," and "teaching"; the titles of its asso- ciated articles confirm that the topic is linked with language pedagogy (fig. 3.6). While some of the decline in articles on language instruction is surely an artifact of the corpus (in 1968 *The German Quarterly* split off

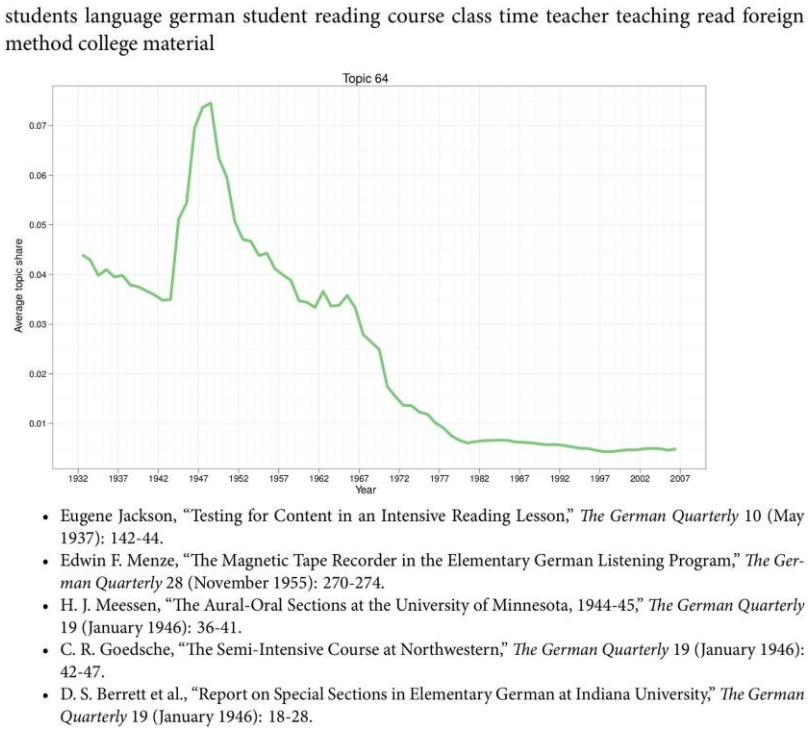


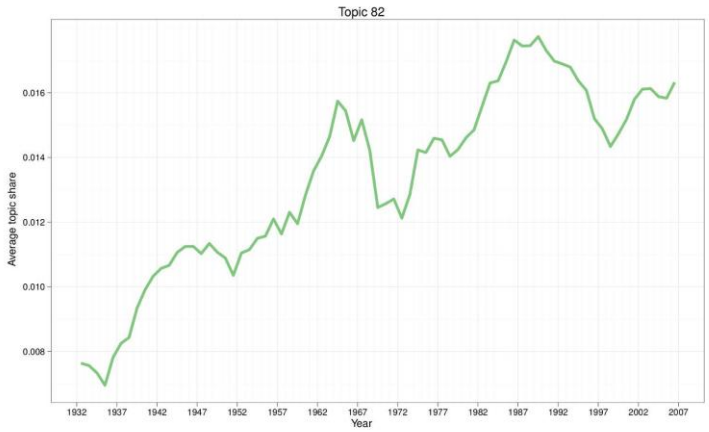
Figure 3.6. Topic 64: Characteristic words, five-year moving average, and repre- sentative articles

a separate journal for language instruction, *Die Unterrichspraxis*, which is not included in the corpus), the decline in the share of these articles is visible well before 1968.

The second trend is the gradual rise in articles concerned with lit- erature and literary criticism (fig. 3.7). This trend is connected with a topic characterized by words such as "literature," "literary," "writers," and "authors."

The recent history of US universities offers a context for these two trends. Both are characteristic of an expansionary period—the "golden age" of higher education in the United States. During this period—roughly between 1945 and 1975—the number of graduate students increased nearly 900 percent. In the 1960s, the number of doctorates awarded every year tripled. The Cold War is often cited among the factors contributing to the expansion of higher education generally and of graduate education in particular. In this period, research displaced teaching as the defining task of the professor. Research for scholars in the humanities was associated with literary history and, eventually, literary criticism.[25]

literature literary german writers authors century writer writing author period book contemporary texts novels



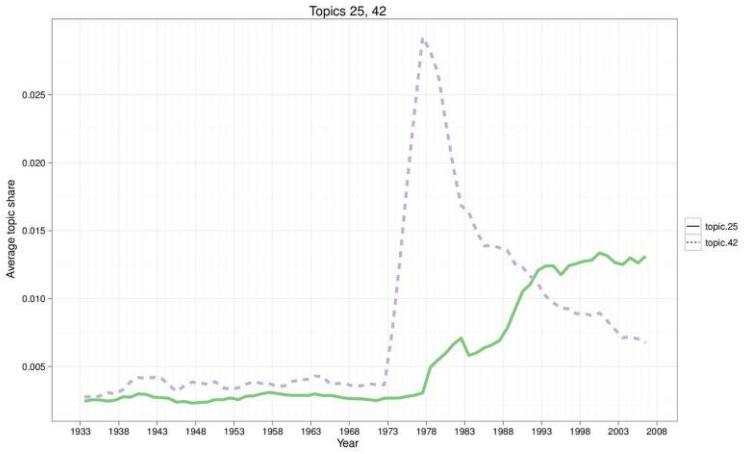
- Leland R. Phelps, review of *The Emergence of German as a Literary Language* by Eric A. Blackall, *Monatshefte* 52 (April-May 1960): 213-14.
- Andreas Kiryakakis, review of *Dictionary of Literary Biography: Volume 66: German Fiction Writers, 1885-1913 Part I: A-L* by James Hardin, *German Studies Review* 13 (May 1990): 331-32.
- Marianne Henn, review of *Benedikte Naubert (1756-1819) and Her Relations to English Culture* by Hilary Brown, *The German Quarterly* 79 (Fall 2006): 532-33.
- Stephen Brockmann, review of *German Literature of the 1990s and Beyond: Normalization and the Berlin Republic* by Stuart Taberner, *Monatshefte* 98 (Summer 2006): 318-19.
- Willa Schmidt, review of *German Fiction Writers, 1885-1913* by James Hardin *Monatshefte* 85 (Spring 1993): 99-101.

Figure 3.7. Topic 82: Characteristic words, five-year moving average, and repre- sentative articles

In addition to the decline of articles on teaching and rise of articles on research, two other topics exhibit distinctive trends (fig. 3.8). The first topic I associate with feminist criticism. Articles connected with this topic appear much more frequently after 1975. The second topic tracks the arrival of the journal *New German Critique* in 1974. Words strongly asso- ciated with the topic include "social," "bourgeois," "political," "class,"

Topic 25: women female woman male feminist gender sexual feminine social role patriarchal movement sex roles masculine

Topic 42: social bourgeois class political critique society theory historical capitalist production marxist marx revolutionary capitalism economic



Topic 25

- Elizabeth Heineman, "Gender Identity in the Wandervogel Movement," *German Studies Review* 12 (May 1989): 249-70.
- Agatha Schwartz, "Austrian Fin-de-Siècle Gender Heteroglossia: The Dialogism of Misogyny, Feminism, and Viriphobia," *German Studies Review* 28 (May 2005): 347-66.
- Maria Dobozy, "Women and Family Life in Early Modern German Literature," *Monatshefte* 98 (Spring 2006): 133-35.
- Meredith Lee, "Der androgyne Mensch: 'Bild' und 'Gestalt' der Frau und des Mannes im Werk Goethes," *The German Quarterly* 71 (Spring 1998): 186-87.
- Ursula Mahlendorf, "Frauen und Gewalt. Interdisziplinäre Untersuchungen zu geschlechtsgebundener Gewalt in Theorie und Praxis," *Monatshefte* 98 (Spring 2006): 141-43.

Topic 42

- Karl Korsch, "The Crisis of Marxism," *New German Critique*, no. 3 (Autumn 1974): 187-207.
- Rainer Paris, "Class Structure and Legitimatory Public Sphere: A Hypothesis on the Continued Existence of Class Relationships and the Problem of Legitimation in Transitional Societies," *New German Critique*, no. 5 (Spring 1975): 149-57.
- Herbert Marcuse, "The Failure of the New Left?" *New German Critique*, no. 18 (Autumn 1979): 3-11.
- Paul Piccone, "Korsch in Spain," review of *Karl Korsch o el Nacimiento de una Nueva Epoca*, ed. Eduardo Subirats, *New German Critique*, no. 6 (Autumn 1975): 148-63.
- Paul Piccone, "From Tragedy to Farce: The Return of Critical Theory," *New German Critique*, no. 7 (Winter 1976): 91-104.

Figure 3.8. Topics 25 and 42: Characteristic words, five-year moving averages, and representative articles

and "society." Herbert Marcuse's "The Failure of the New Left" numbers among the articles most strongly associated with this topic. None of the words comes as a surprise to those familiar with the journal. Its publisher describes the journal as having "played a significant role in introducing US readers to Frankfurt School thinkers."[26]

All the topics mentioned so far appear in different proportions in the corpus. Figure 3.9 shows the frequency of several topics over time on the same scale. Recall that what is being counted on the vertical axis is the average topic share among all articles in a given year (or the average proportion of all words in a given year associated with a given topic). If we accept for a moment the analogy between subject matter and topic, it would mean that a year with ten articles published and a 0.1 average share for the topic associated with language pedagogy might have two articles with half their words associated with the pedagogy topic. Or it might be the case that for all ten articles, one-tenth of their words were associated with the pedagogy topic. In either case, the average topic share is 0.1. It is also worth emphasizing that the LDA model makes use of relative rather than absolute word frequencies. That is, a 500-word review that is 20 percent topic 64 is treated the same, in certain important respects, as a 9,000-word article that is 20 percent topic 64, even though the number of words and share of space in the journal are different. Infrequent top- ics also bring with them their own set of concerns. With topics associated with only a few articles a year, such as the "folktales" topic discussed later, selection bias becomes a concern. It is possible that some trends are not

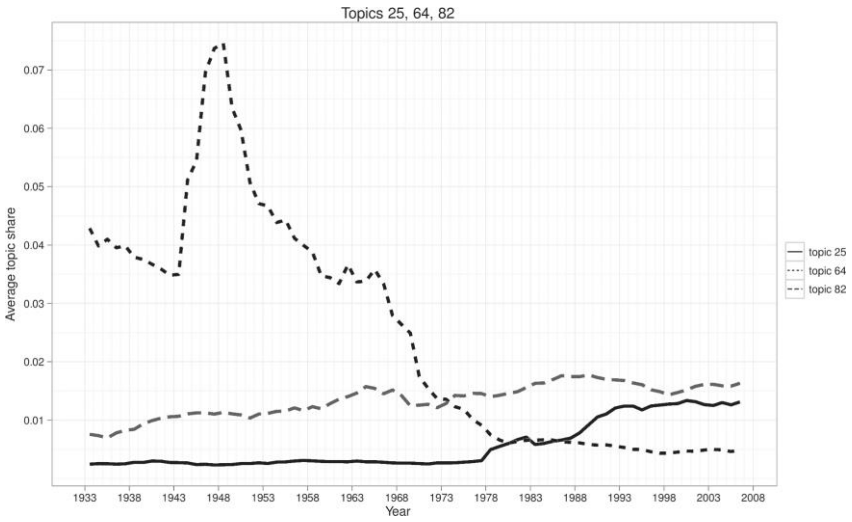


Figure 3.9. Comparison of topics 25 ("women . . ."), 64 ("students . . ."), and 82 ("literature . . .")

real in the sense that a rapid decline might reflect a certain kind of article migrating elsewhere—perhaps to a European history journal—rather than any decline in research on the subject in German studies generally.

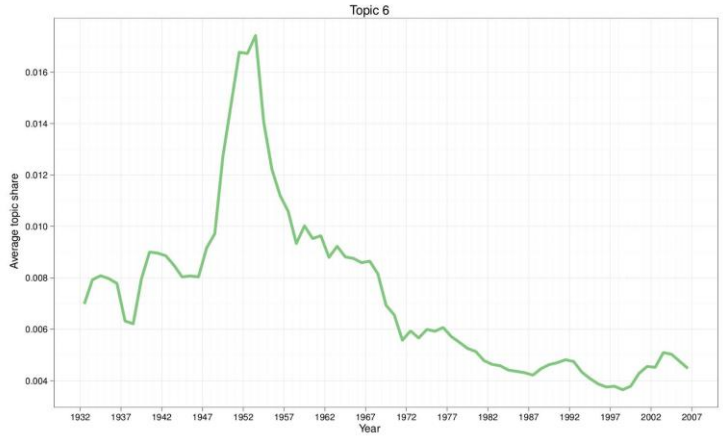
Long Nineteenth-Century Topics

Two topics that track specific areas of nineteenth-century scholarship are worth mentioning, as their trajectory over the period reveals predictable rhythms of scholarly publishing.

A single topic is associated with articles on the life and works of Goethe (fig. 3.10). A rapid increase in articles associated with this topic begins around 1947. This surge of articles coincides with the bicentennial of Goethe's birth (1749). *The German Quarterly*, for example, devoted the entire November 1949 issue to the bicentennial. That the topic model reflects this as well as it does offers additional validation that it is capable of capturing the gross features of the corpus.

Another topic identifies scholarship connected to folktales (fig. 3.11). With peaks around 1955 and 1990, there is a temptation to think that

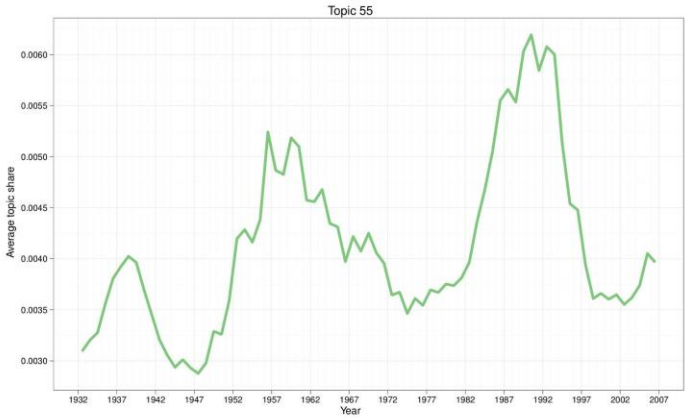
goethe faust goethes wilhelm werther weimar iphigenie ottilie gretchen charlotte meisters
mephisto meister dichtung wahlverwandtschaften



- L. M. Price, "Goethe Bibliography for 1939," *Monatshefte für deutschen Unterricht* 32, no. 2 (February 1940):83-88.
- Heinz Bluhm, "Goethe Bibliography for 1942 to 1944: German Non-Periodical Publications," *Monatshefte* 39, no. 2 (February 1947): 126-33.
- J. A. Kelly, "Goethe Bibliography for 1938," *Monatshefte für deutschen Unterricht* 31, no. 8 (December 1939): 400-06.
- Heinz Moenkemeyer, "Zum Verhältnis von Sorge, Furcht und Hoffnung in Goethes Faust," *The German Quarterly* 32, no. 2 (March 1959): 121-32.
- Hellmut Ammerlahn, "Mignons nachgetragene Vorgeschichte und das Inzestmotiv: Zur Genese und Symbolik der Goetheschen Geniusgestalten," *Monatshefte* 64, no. 1 (Spring 1972): 15-24.

Figure 3.10. Topic 6: Characteristic words, five-year moving average, and repre- sentative articles

tale tales fairy grimm folk wilhelm stories jacob brothers tradition grimms folklore magic story popular



- Maria M. Tatar, review of Breaking the Magic Spell: Radical Theories of Folk and Fairy Tales by Jack Zipes, *The German Quarterly* 55, no. 2 (March 1982): 231-32.
- Ruth B. Bottigheimer, review of One Fairy Story Too Many: The Brothers Grimm and Their Tales by John M. Ellis, Fairy Tales and the Art of Subversion: The Classical Genre for Children and the Process of Civilization by Jack Zipes, The Trials and Tribulations of Little Red Riding Hood: Versions of the Tale in Sociocultural Context by Jack Zipes, and Die Geschichte vom Rotkäppchen: Ursprünge, Analysen, Parodien eines Märchens by Hans Ritz, *The German Quarterly* 58, no. 1 (Winter 1985): 144-47.
- Ruth B. Bottigheimer, "Sixteenth-Century Tale Collections and Their Use in the 'Kinder- und Hausmärchen,'" *Monatshefte* 82, no. 4 (Winter 1992): 472-90.
- Ruth B. Bottigheimer, "Tale Spinners: Submerged Voices in Grimms' Fairy Tales," *New German Critique*, no. 27 (Autumn 1982): 141-50.
- Donald P. Haase, review of The Trials and Tribulations of Little Red Riding Hood: Versions of the Tale in Sociocultural Context by Jack Zipes, *Monatshefte* 78, no. 3 (Fall 1986): 385-86.

Figure 3.11. Topic 55: Characteristic words, five-year moving average, and repre- sentative articles

interest in folktales may rise and fall in a regular cycle. Yet further reflec- tion yields a simpler explanation for the second rise: the anniversary of the births of Jacob and Wilhelm Grimm (1785 and 1786, respectively). The fluctuations in the topic's prevalence before 1970 may be due to a num- ber of factors. For example, the arrival of new journals emphasizing schol- arship on twentieth-century subjects seems likely to have contributed to the decline in the relative share of articles concerned with scholarship on folktales.

Topic-Modeling Pitfalls

While LDA has proven an effective method for exploring very large collections of texts, it has important shortcomings, some of which are shared by other topic models. First, topics lack an interpretation apart from the probabilistic model in use. Articles may be compared in terms of their topics—one such measurement is called the Kullbeck-Leibler divergence—but this metric suffers from problems of interpretation

familiar from the discussion of cosine distance. Moreover, recent work has shown that automatic measures of the fit between a topic model and a corpus (e.g., held-out likelihood) do not always align with human readers' assessments of the coherence of inferred topics, suggesting a mismatch at some level between topic models and topics familiar to human read- ers.[27] Given this shortcoming, it becomes essential that those using topic models validate the description provided by a topic model by reference to something other than the topic model itself. Fortunately researchers familiar with the period, documents, and writers associated with a corpus typically have the expertise to devise appropriate checks.

An additional complication is the fact that the number of topics in a model is *arbitrary*. In this chapter, I made use of a thirty-topic fit (fig. 3.5) and a hundred-topic fit to characterize the same corpus of jour- nal articles. While many of the topics of the thirty-topic fit resemble those of the hundred-topic fit, the topics are distinct. That the number of topics and the composition of the inferred topics can vary in this manner should reinforce the idea that an individual topic has no interpretation outside the particular model in use. Blei and his coauthors are admirably clear on this point.[28]

LDA and other topic models also make assumptions known to be incorrect.[29] For example, LDA assumes that the association of words with a topic does not vary over time. In other words, LDA assumes scholars are using the *same collection of words* to talk about folktales in the year 1940 and the year 2000. We know this is wrong. That LDA works as well as it does is due to the fact that many words are used consistently over time. That is, regardless of the decade in which the articles were written, articles about Goethe's life will tend to use words like "Goethe" and "Faust." For other kinds of inquiry, especially those concerned with less conspicu- ous trends, changes in language use are a significant concern. Changes in terminology in particular—for example, if writers systematically begin using "folklore" in a context where they previously would have used "folktales"—present a potential problem for LDA. For all these reasons, the assumptions made by topic models require close and careful reading.

Prospects for Topic Models

Long nineteenth-century materials, in particular, are unusually hospita- ble to the use of machine reading and probabilistic models. A staggering amount of printed material survives to the present day. Moreover, these texts are all unencumbered by copyright in the United States. Contrast this with the disposition of materials published in the twentieth century. Scholars working with printed material from the twentieth century are hamstrung by copyright law—unable to share text collections freely if the collections contain works published after 1924.

For researchers in the humanities and interpretive social sciences, learning how to use and reflect critically about models such as LDA is growing easier. Leading universities such as MIT and Stanford have announced a number of freely accessible online courses that cover prob- ability and computational linguistics. These courses discuss the bag-of- words model and probabilistic models of text collections. One such course is taught by Andrew Ng, the third author of the original LDA paper.

This chapter has made no attempt to use topic models to investi- gate existing accounts of the history of German studies. Beginning with specific hypotheses, however, often makes for compelling research. Per- haps unsurprisingly, it has been computational linguists who have pio- neered using topic models to ask specific questions about the history of their own discipline.[30] For example, David Hall takes up a hypothesis inspired by Thomas Kuhn's account of the historical trajectory of sci- ence as one punctuated by periodic "revolutions" in dominant meth- ods.[31] Hall observes that there have been widely acknowledged shifts in the prominence of certain methods within computational linguistics over the past twenty years. If these methodological shifts represented a revolutionary change of "paradigm" in Kuhn's sense, then Hall antici- pated that the researchers associated with "insurgent" methods would not be participants in a field—that is, authors of articles—with long stand- ing. In other words, these researchers would be new arrivals, not estab- lished scholars abandoning existing methodologies in favor of new ones. A topic model of journal articles allowed Hall to identify significant methodological shifts in the discipline and those authors associated with the changes. This general line of inquiry—with or without the guiding Kuhnian perspective—could be adapted to a number of other disciplines, including German studies. As this chapter has demonstrated, there are a number of changes in method and subject matter that are visible in the discipline's journals since 1928. Future research might use quantitative methods to identify the scholars associated with these shifts.

My aim in this chapter has been to show that a topic model reveals disciplinary trends that would otherwise be prohibitively time consum- ing to document. Used alongside direct and collaborative reading, topic models have the potential to offer new perspectives on existing materials and novel accounts of the dynamics of intellectual history.

Notes

1 Sharon Block and David Newman, "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts," *Journal of Women's History* 23, no. 1 (2011): 81–109; Justin Grimmer, "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis* 18, no. 1 (2010): 1–35; David Hall, "Tracking the

Evolution of Science" (bachelor's thesis, Stanford University, 2008); David Hall, Daniel Jurafsky, and Christopher D. Manning, "Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI: Association for Computational Lin- guistics, 2008), 363–71; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *ACM Journal of Computing in Cul- tural Heritage* 5, no. 1 (2012), doi:10.1145/2160165.2160168.

2 Gregory Crane, "What Do You Do with a Million Books?" *D-Lib Magazine* 12, no. 3 (March 2006), doi:10.1045/march2006-crane.

3 David Mimno and David Blei, "Bayesian Checking for Topic Models," in *Pro- ceedings of the 2011 Conference on Empirical Methods in Natural Language Pro- cessing* (Somerset, NJ: Association for Computational Linguistics, 2011), 227–37; Robert K. Nelson, "Mining the Dispatch," Digital Scholarship Lab, University of Richmond, accessed March 28, 2012, http://dsl.richmond.edu/dispatch.

4 Laurel Ulrich, *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary, 1785–1812* (New York, NY: Knopf, 1990).

5 Kirsten Belgum, *Popularizing the Nation: Audience, Representation, and the Production of Identity in Die Gartenlaube, 1853–1900* (Lincoln: University of Nebraska Press, 1998); Fritz K. Ringer, *The Decline of the German Mandarins: The German Academic Community, 1890–1933* (Cambridge, MA: Harvard Uni- versity Press, 1969).

6 Larry Isaac, "Movements, Aesthetics, and Markets in Literary Change: Mak- ing the American Labor Problem Novel," *American Sociological Review* 74, no. 6 (2009): 938–65, doi:10.1177/000312240907400605; Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005); Carl P. Simon and Eric S. Rabkin, "Culture, Science Fiction, and Complex Adaptive Sys- tems: The Work of the Genre Evolution Project," in *Biocomplexity at the Cutting Edge of Physics, Systems Biology and Humanities*, ed. Gastone Castellani et al. (Bolo- gna: Bononia University Press, 2008), 279–94; John Unsworth, "20th-Century American Bestsellers," accessed October 29, 2013, http://people.lis.illinois.edu/~unsworth/courses/bestsellers.

7 Eric S. Rabkin, "Science Fiction and the Future of Criticism," *PMLA* 119, no. 3 (2004): 457–73; Simon and Rabkin, "Culture, Science Fiction, and Complex Adap- tive Systems."

8 Isaac, "Movements, Aesthetics, and Markets in Literary Change."

9 N. Katherine Hayles, *How We Think: Digital Media and Contemporary Techno- genesis* (Chicago: University of Chicago Press, 2012), 55–80.

10 *Monatshefte* changed its name three times between 1899 and 1946. While referred to simply as *Monatshefte* in the United States, its full title since 1946 has been *Monatshefte für deutschsprachige Literatur und Kultur*. The original size of the corpus provided by JSTOR was 26,104 documents. From this initial cor- pus, I removed articles flagged by JSTOR as "misc," typically front matter and advertisements, as well as documents having fewer than two hundred words. This yielded the corpus of 22,198. To facilitate computation, rare words (those occur- ring in fewer than ten documents) were removed, along with extremely frequent

words in German and English (so-called stop words) and words with only one or two characters. The size of the remaining lexicon was 74,158 unique terms. The total number of words in all articles was 15,680,621.

11 This final step—removing all numbers—creates a special problem with this cor- pus. Since the Eszett (ß) is mangled by JSTOR OCR into "l3," all words contain- ing ß are removed as they contain a numeric character ("3"). Given the nature of this present inquiry—the concern for clear trends visible across many articles—this does not present a serious problem: any easily detectable trend in the corpus will be the product of *many* words systematically co-occurring.

12 James Boyle, *The Public Domain: Enclosing the Commons of the Mind* (New Haven, CT: Yale University Press, 2008); Lawrence Lessig, *Free Culture: The Nature and Future of Creativity* (New York: Penguin Press, 2005).

13 Formally, we might consider a bag in the context of the following three concepts: set, bag, and sequence. A set is an unordered list of elements that ignores order and duplicates, $S = \{4,4,5\} = \{4,5\}$. A bag is an unordered list that takes into account repeated elements, $B = \{4,4,4,5\} = \{5,4,4,4\}$. A sequence considers both order and repeated elements, $Q = \{4,4,5\} \neq \{5,4,4\}$.

14 Michael J. Crowe, *A History of Vector Analysis: The Evolution of the Idea of a Vectorial System* (Notre Dame, IN: University of Notre Dame Press, 1967).

15 Christopher D. Manning and Hinrich Schüzte, *Foundations of Statistical Natu- ral Language Processing* (Cambridge, MA: MIT Press, 1999).

16 Michael Lee, Brandon Pincombe, and Matthew Welsh, "An Empirical Evalu- ation of Models of Text Document Similarity," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Erlbaum, 2005), 1254–59.

17 David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Alloca- tion," *Journal of Machine Learning Research* 3 (2003): 993–1022.

18 Dirichlet was a contemporary of Carl Friedrich Gauss and Carl Gustav Jacobi. Alexander von Humbolt supported his candidacy to the Prussian Academy of Sci- ences. Through Humbolt he met his future wife, Rebecka Mendelssohn, sister of the composer Felix Mendelssohn and granddaughter of Moses Mendelssohn. Dirichlet played a vital role in the development of modern mathematics, the mod- ern definition of a function being credited to him. See I. M. James, *Remarkable Mathematicians: From Euler to von Neumann* (Washington, DC: Mathematical Association of America, 2002).

19 David Blei, "Introduction to Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (2012): 77–84, doi:10.1145/2133806.2133826. Blei's
commentary is worth repeating: "Indeed calling these models 'topic models' is retrospective—the topics that emerge from the inference algorithm are interpre- table for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA" (79).

20 For subsequent developments, see David M. Blei and John D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, ed. William Cohen and Andrew Moore (Pittsburgh, PA: